State of the Cephalopod

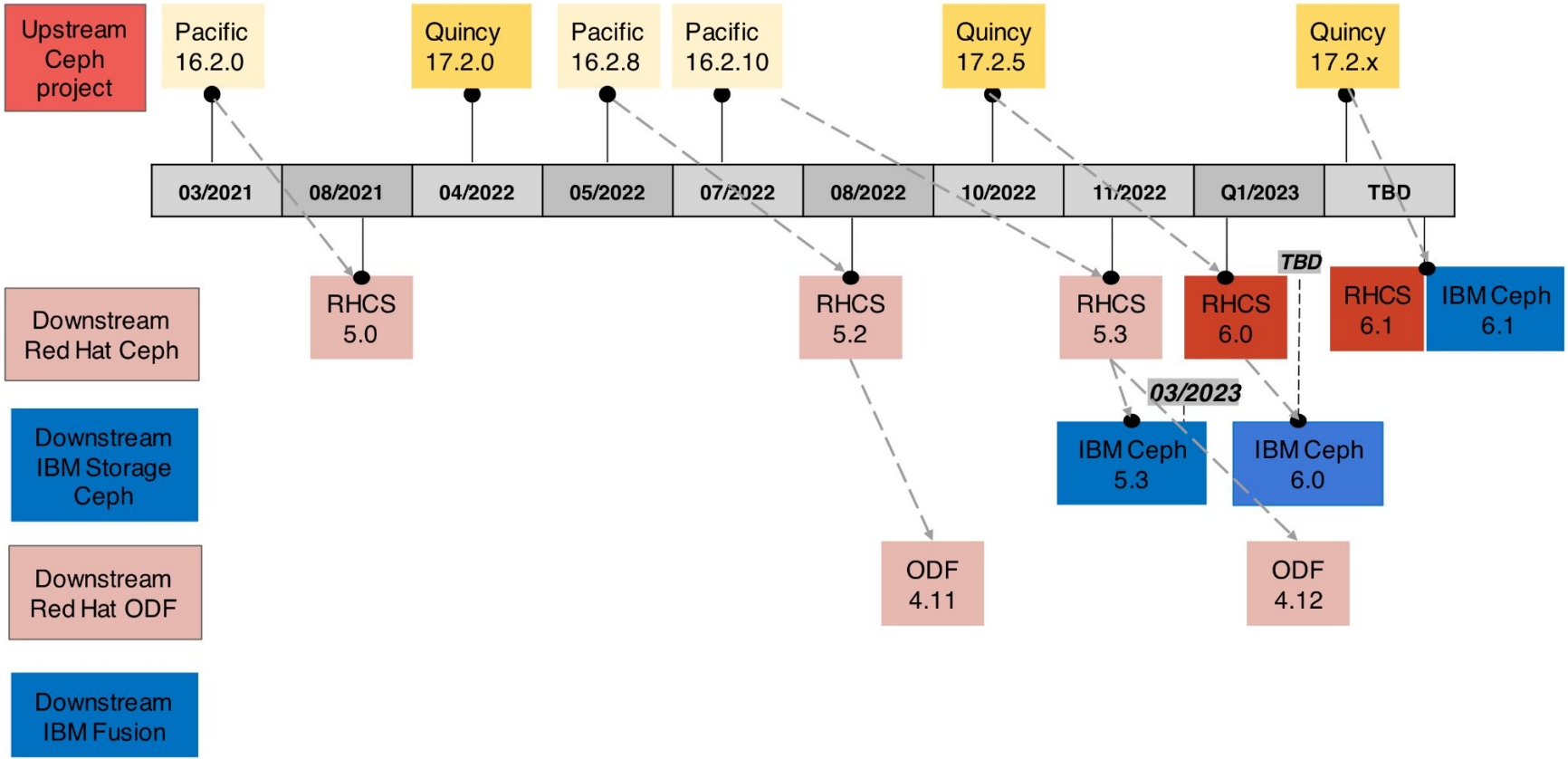
2023.06.14

Federico Lucifredi

Josh Durgin


From Red Hat to IBM

DOWNSTREAM TIMELINES



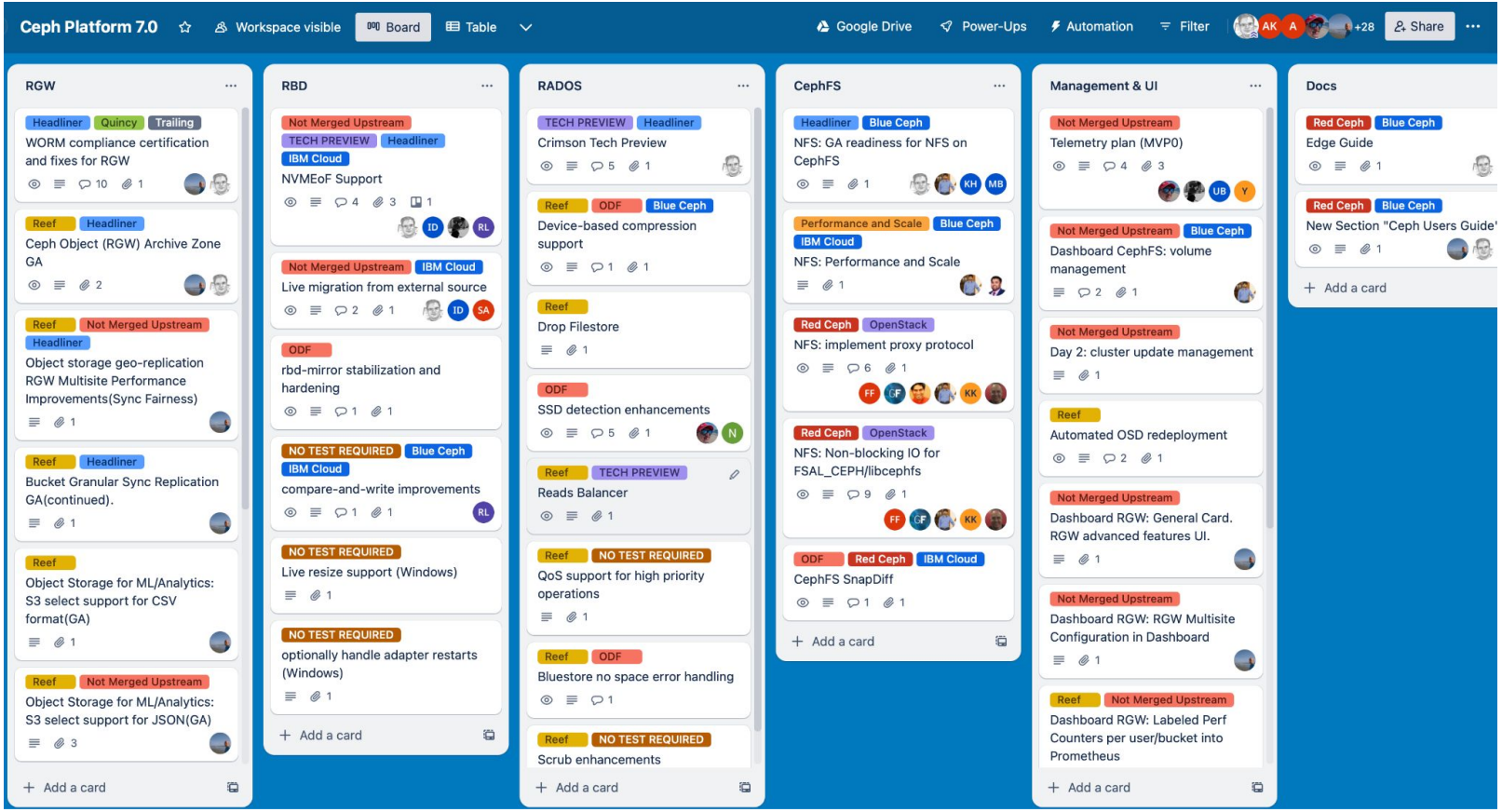
CEPH AS A PRODUCT PLATFORM

- Red Hat Ceph Storage
  - Red Hat OpenStack Platform
  - Object petabyte-scale clusters
- OpenShift Data Foundation
  - IBM Fusion
- IBM Storage Ceph
  - Object petabyte-scale clusters
  - Unified Storage (soon!)
- IBM Cloud
  - Storage technology
- Appliance

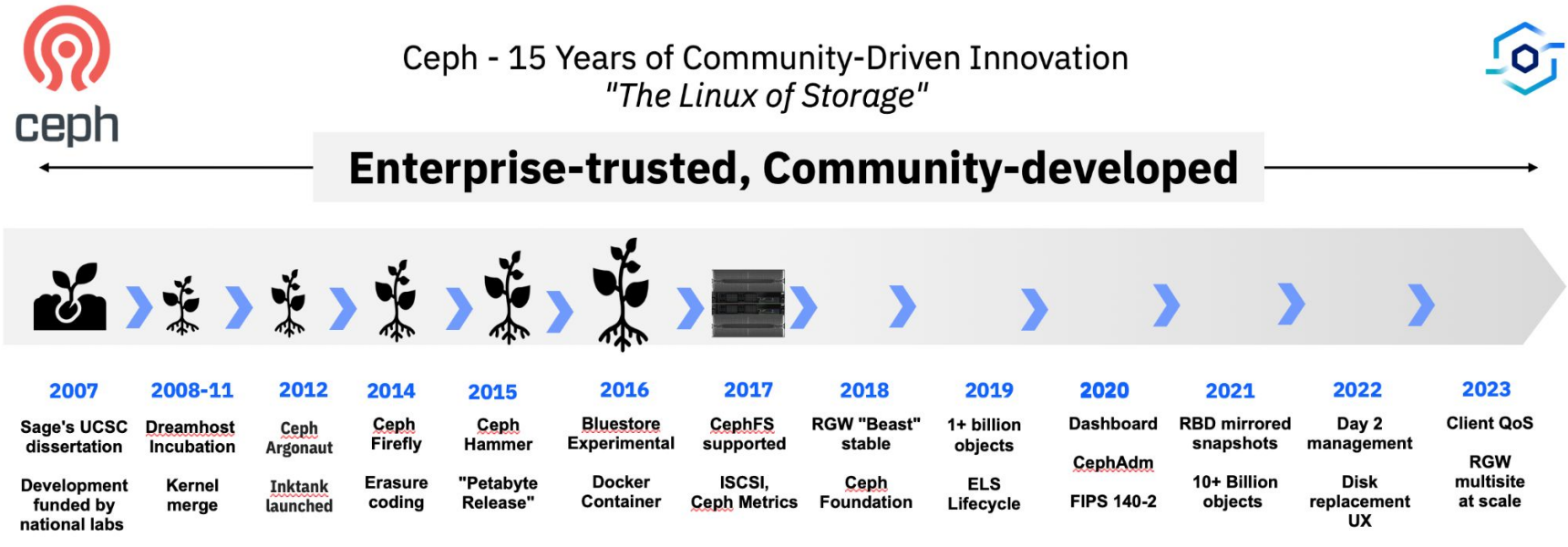
CEPH PLATFORM

- **One platform, many products**
  - **IBM Storage Ceph or Red Hat Ceph Storage? Same code**
  - **Different commercial choices**
- **One team, many products**
  - The same folks are behind all of them
- **Three-year lifecycle, five year extension**
  - CVEs, hot-fixes, bug-fixes and all that
  - Support, of course
- Upgrade Paths
  - Red-to-blue “crossgrade” (and Community)
  - N+2 upgrades

WHERE NEXT


Ceph Platform 7.0
Workspace visible
Board
Table
Google Drive
Power-Ups
Automation
Filter
+28
Share
RGW
Headliner Quincy Trailing
WORM compliance certification and fixes for RGW
Reef Headliner
Ceph Object (RGW) Archive Zone GA
Reef Not Merged Upstream Headliner
Object storage geo-replication RGW Multisite Performance Improvements(Sync Fairness)
Reef Headliner
Bucket Granular Sync Replication GA(continued).
Reef
Object Storage for ML/Analytics: S3 select support for CSV format(GA)
Reef Not Merged Upstream
Object Storage for ML/Analytics: S3 select support for JSON(GA)
Add a card
RBD
Not Merged Upstream TECH PREVIEW Headliner IBM Cloud
NVMEoF Support
Not Merged Upstream IBM Cloud
Live migration from external source
ODF
rbd-mirror stabilization and hardening
NO TEST REQUIRED Blue Ceph IBM Cloud
compare-and-write improvements
NO TEST REQUIRED
Live resize support (Windows)
NO TEST REQUIRED
optionally handle adapter restarts (Windows)
Add a card
RADOS
TECH PREVIEW Headliner
Crimson Tech Preview
Reef ODF Blue Ceph
Device-based compression support
Reef
Drop Filestore
ODF
SSD detection enhancements
Reef TECH PREVIEW
Reads Balancer
Reef NO TEST REQUIRED
QoS support for high priority operations
Reef ODF
Bluestore no space error handling
Reef NO TEST REQUIRED
Scrub enhancements
Add a card
CephFS
Headliner Blue Ceph
NFS: GA readiness for NFS on CephFS
Performance and Scale Blue Ceph IBM Cloud
NFS: Performance and Scale
Red Ceph OpenStack
NFS: implement proxy protocol
Red Ceph OpenStack
NFS: Non-blocking IO for FSAL_CEPH/libcephfs
ODF Red Ceph IBM Cloud
CephFS SnapDiff
Add a card
Management & UI
Not Merged Upstream
Telemetry plan (MVP0)
Not Merged Upstream Blue Ceph
Dashboard CephFS: volume management
Not Merged Upstream
Day 2: cluster update management
Reef
Automated OSD redeployment
Not Merged Upstream
Dashboard RGW: General Card. RGW advanced features UI.
Not Merged Upstream
Dashboard RGW: RGW Multisite Configuration in Dashboard
Reef Not Merged Upstream
Dashboard RGW: Labeled Perf Counters per user/bucket into Prometheus
Add a card
Docs
Red Ceph Blue Ceph
Edge Guide
Red Ceph Blue Ceph
New Section "Ceph Users Guide"
Add a card

JUST BEGINNING



Vibrant open source developer community

- 1000+ contributors
- 200+ organizations
- 600K+ lines of code changed
- 17,000 code commits

Vibrant open source user community

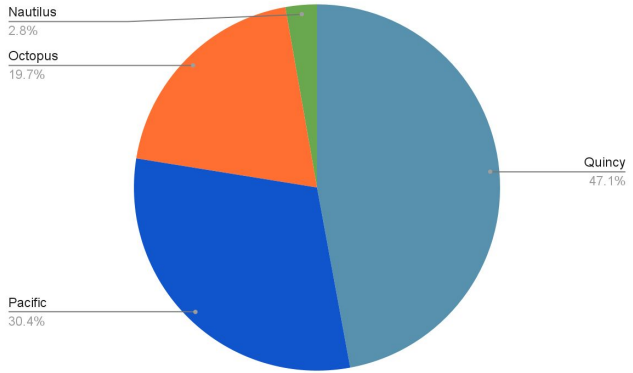
- 3 to 5 EB deployed, examples include:
  - CERN
    NASA
    Flipkart
    Salesforce

Project Update

PROJECT UPDATE

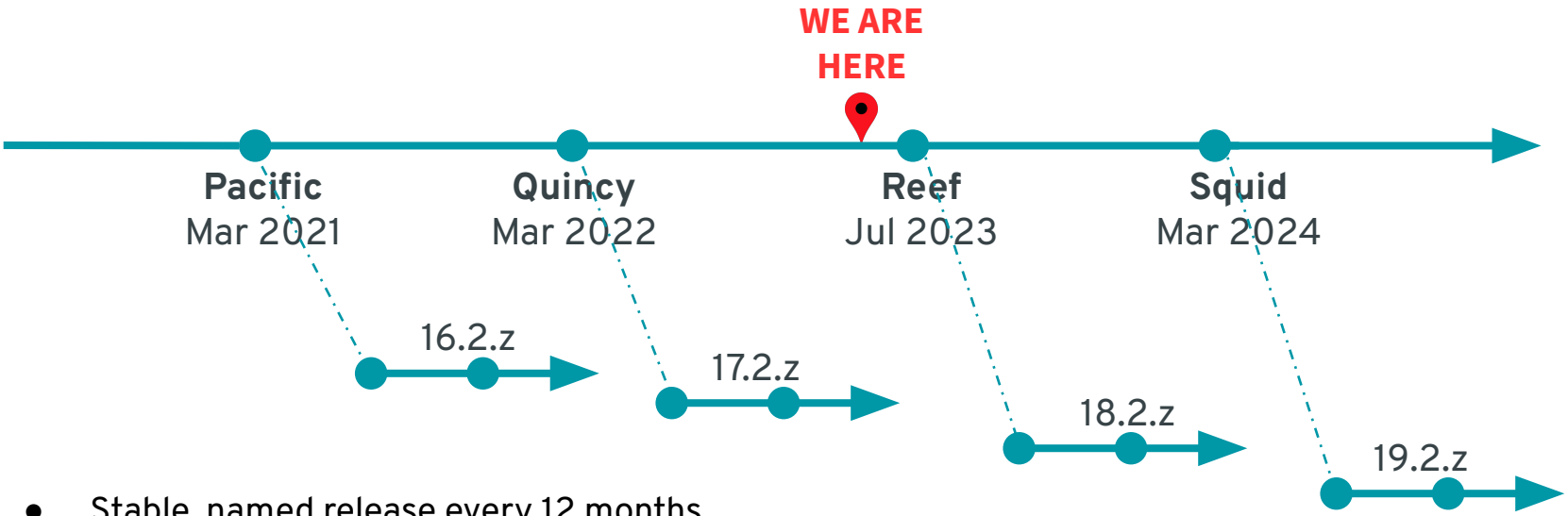
- New Ceph Governance Model
  - 3-member elected Executive Council
    - Project coordination, single contact point
    - Interface with the foundation
  - Driven by the Ceph Leadership Team - group effort, shared leadership, open meetings
- Recent Focus Areas
  - Release process
    - Publishing RC candidates
    - Multiple real-world upgrades before release
  - Performance and scalability hardening
    - Pawsey, other major scale tests with 1000s of OSDs
    - Logical large scale tests in teuthology

- Telemetry data:
  - About 2.5K reporting clusters
  - >1 Exabyte total capacity, >150K OSDs



- Public dashboards:
  - telemetry-public.ceph.com

RELEASE SCHEDULE



- Stable, named release every 12 months
- Backports for 2 releases
  - Pacific reaches EOL shortly after Reef is released
- Upgrade up to 2 releases at a time
  - Nautilus → Pacific, Pacific → Reef, Quincy → Squid
- Client compatibility 3 releases back
  - Pacific clients will be able to talk to Squid clusters


Reef and Beyond

PERFORMANCE & SCALE

- Pawsey Supercomputing Centre
  - 4000 OSDs - real world setup
  - 64PB raw capacity
  - Several bottlenecks in cephadm/dashboard/mgr fixed
- Gibba - upstream sepia lab
  - 1000 OSDs - logical scale (limited per-OSD resources)
  - Verifying Quincy, Reef work at scale, e.g. upgrades and background QoS
- Red Hat scale lab
  - 8000 OSDs - logical scale (limited per-OSD resources)
  - Exploring monitoring bottlenecks and solutions
- Performance CI

RADOS - REEF

- Background QoS - improved cost model with extensive testing
- Work started on client vs client QoS
  - Initial implementation in librados and testing
- Support for high priority operations with mClock, eg. forced recovery
- BlueStore
  - custom WAL for RocksDB - WIP
  - Elastic shared blobs logic
  - 4K allocation unit for bluefs, expandable superblocks
- Balancer - workload (primary) balancer
  - Optimizes for balance of reads, in addition to writes
- PG log improvements to avoid and detect memory growth
- Stretch cluster bug fixes and test expansion

CRIMSON PROJECT

- High-performance rewrite of the OSD, currently supports RBD workloads on replicated pools with BlueStore
- New in Reef:
  - Usability improvements (set-allow-crimson, crimson pool type)
  - Snapshot support
  - Essentially complete rados api coverage with teuthology testing (for replicated pools)
- Planned for Squid:
  - Scrub
  - Performance!
  - Messenger and other multi-reactor improvements

CRIMSON PROJECT - SEASTORE

- Objectstore implementation for Crimson
- New in Reef:
  - Support for rbd workloads
  - SeaStore metrics
  - Initial tiering architecture
- Planned for Squid:
  - LBA tree traversal optimizations
  - ObjectDataHandler support for reading and writing sub-extents
  - Support for promoting hot extents on read to fast tier
  - Circular journal and in-place writes for fast media
  - Multi-reactor support
  - laddr redirection support for CoW snapshots

TELEMETRY - REEF

- Work continues on backend analysis of telemetry data
  - Tools for developers to use crash reports to identify and prioritize bug fixes
  - Perf counters analysis
- Adjustments in collected data
  - Adjust what data is collected for Reef
  - Periodic backport to Quincy (we re-opt-in)
  - e.g., which orchestrator module is in use (if any)
- Upgraded telemetry server
  - Allows for faster query execution
- Drive failure prediction
  - Building improved models for predictive drive failures
    - Collaborating with drive manufacturers

TELEMETRY - SQUID

- Work continues on backend analysis of telemetry data
- Adjustments in collected data
- Drive failure prediction
  - Continuing collaboration with drive manufacturers to improve models for predictive drive failures
- In depth cluster x-ray view
  - To allow for crash tracking and time series insights
- Periodic newsletters

DASHBOARD - REEF

- New and improved Landing Page
- RGW Server-side encryption
- Complete support for RBD (RBD Mirroring)
- Operational improvements
  - 1-click OSD creation
  - Improved capacity planning
  - Ceph auth and user listing
- Observability
  - Centralized Logging (Grafana-Loki based)
- Accessibility
  - Compliant with WCAG level AA.

DASHBOARD - SQUID

- RGW Advanced Workflows (user roles/policies, bucket policies, lifecycle, notifications...)
- RGW Multi-site Workflow
- Support for NVMe Management
- Support for CephFS Management
- Cluster Upgrades
- Continuous UI/UX Improvements
- Replacing Grafana with built-in charts

RGW

- S3 Select Enhancements
  - Json object format support in Reef
  - Trino integration in progress
- S3 Inventory
- Multisite Performance and Scalability
  - Sync fairness - load balancing across RGWs
  - Testing and stabilization of per-bucket replication
- HTTP/3 Frontend Prototype
- RGW Standalone Prototype
  - Posix-based file backend

CEPHADM - REEF

- OS Tuning Profiles
  - Manage sysctl settings across hosts using cephadm
- Staggered Upgrades
  - Allow upgrading by one daemon type/service at a time
  - Can tell cephadm to only upgrade X number of daemons then stop
- Simplified rgw multisite workflow
  - Still WIP, should be done for Reef release
- Cephadm is now “compiled” (by py zipapp)
  - Will allow splitting the (nearly 10000 line) cephadm binary into multiple files.
  - Should have minimal user impact
  - Will be publishing the “compiled” version with the release instead of expecting users to curl from github
  - Should also be simple for users to “compile” on their own from the source tree as long as they have Python >= 3.5 (just run the “build.py” python script)
- Auth Key rotation for ceph daemons
  - ceph orch daemon rotate-key <daemon-name>

ROOK

- Rook v1.11
  - Supports Pacific and Quincy
- Rook v1.12
  - Tentatively in July
- Planned support for Reef
  - Either for v1.11, or v1.12 depending on Reef timing
- Recent features
  - Support for the Ceph exporter daemon
  - Mirroring across clusters with overlapping networks, based on multi-cluster services
    - Globalnet Submariner
  - OSD encryption key rotation
  - Bucket notifications and topics declared stable

RBD

- Encryption-formatted copy-on-write clones
  - Clone images encrypted with encryption format or key different from parent
    - E.g. encrypted clone (key A) of encrypted clone (key B) of unencrypted golden image
  - Enablement support is coming in QEMU 8.0 and libvirt 9.3
- Persistent write-back cache usability (status reporting, etc)
- rbd-mirror stabilization and hardening
  - Ensure correct operation when daemon restarts - pick up where it left off
  - Consistent per-image metrics (using new per-node exporter framework)
  - Ongoing scale testing
- NVMeoF target gateway
  - Initial single-gateway-in-single-gateway-group implementation
  - Discovery service, deployment implementation in progress
- Research into log-structured data format - https://github.com/ceph/ceph/pull/49549

CEPHFS

- continued improvements for fscrypt support (*)
- much-improved cephfs-top
- snapshot diff support (*)
- more tests on more scenarios
  - workload tests with cephfs subvolumes
- widespread bug fixes, stability, and admin UX improvements
  - Driven by a big increase in number of live deployments
- rebalance subtree to a subset of active MDSs
  - mds_bal_rank_mask
  - use static pins and dynamic subtree balancing without interference!

(*) changes will be backported

TESTING

- Developer experience - quicker feedback loop with local integration tests
  https://github.com/zmc/ceph-devstack/
  - Next step - quick local builds
  - Expand to make running tests in any lab as simple as possible
- Lab infrastructure improvements
  - Large outage this year, open group to improve reliability of the test lab and manage infrastructure
- Engage with more groups to work on scale and correctness testing
- Performance CI - catch regressions as early as possible

DOCUMENTATION

The Team

- three people who, perhaps unusually for software, *really* care about the documentation and are capable of arguing productively and working together to improve it.

Raw Numbers

- over 1800 docs-related commits in the past year
- 1200-1500 lines edited per week, each week, in 2023
- most of these changes can be seen here: https://github.com/zdover23

Send all documentation requests and complaints to Zac Dover <zac.dover@proton.me>

DOCUMENTATION

Things Done

- We developed a reliable, unambiguous workflow for making changes to the documentation (quickly if necessary) and for ensuring that those changes are properly backported.
- We edited the following documents to improve their semantics, syntactics, and pragmatics: cephadm, RGW, crushtool, RADOS, cephfs, Developer Guide, (and many others over the past three years).
- We rectified and expanded the glossary: https://docs.ceph.com/en/latest/glossary/

Things To Be Done

- Write overview articles: outline the most important facts for each subject; suggest further reading.
- Write a Beginners' Guide: create a conceptual overview and maybe a reference architecture that allows people ignorant of but curious about Ceph to understand what Ceph offers.
- Improve Contextual Help (IcePic's request, and a good one)

Send all documentation requests and complaints to Zac Dover <zac.dover@proton.me>

Ceph Day Tomorrow

Room 18

10:20-18:00

More intro and advanced talks

Ceph BoF at 5:30